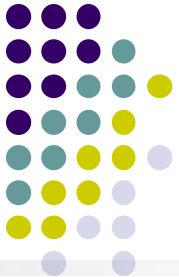
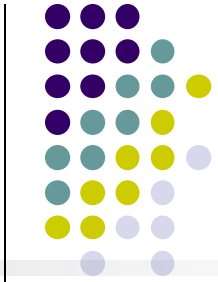


10- 701 Fall 2011
Recitation - Probability Review

Suyash Shringarpure

09/13/11

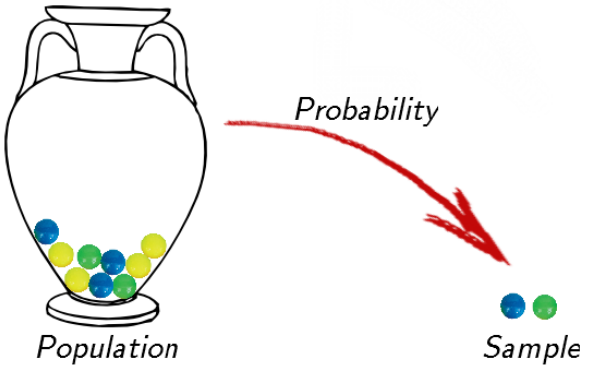
What we will cover



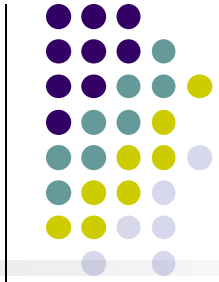
- Basic probability
  - Definitions and Axioms.
  - Random Variables – PDF and CDF.
- Joint distributions.
- Some common distributions.
- Independence.
- Conditional distributions.
- Information theory basics
  - Application to decision trees
- Overfitting and pruning

Probability

- Real world - Full of uncertainty..
- Eg. I have to reach home by 7:30 pm. Can I take the 7:15 pm 61 C at CMU and reach?
  - How much time will the bus take after I get it (possible delays due to traffic, roads, etc)
  - What if the bus arrives late?
- Probability – A mechanism for decision making in the presence of uncertainty
- Probability is a way of using information about a population to learn about a sample.



Why use probability?



- There have been attempts to develop different methodologies for uncertainty:
  - Fuzzy logic
  - Qualitative reasoning (Qualitative physics)
  - …
- In 1931, de Finetti proved that :
  - If you gamble using probability you can't be unfairly exploited by an opponent using some other system

Basic Concepts

- A *sample space* S is the set of all possible outcomes of a conceptual or physical, repeatable experiment. (S can be finite or infinite.)
  - E.g., S may be the set of all possible outcomes of a dice roll:
- An event A is any subset of S.
  - Eg., A= Event that the dice roll is < 3.

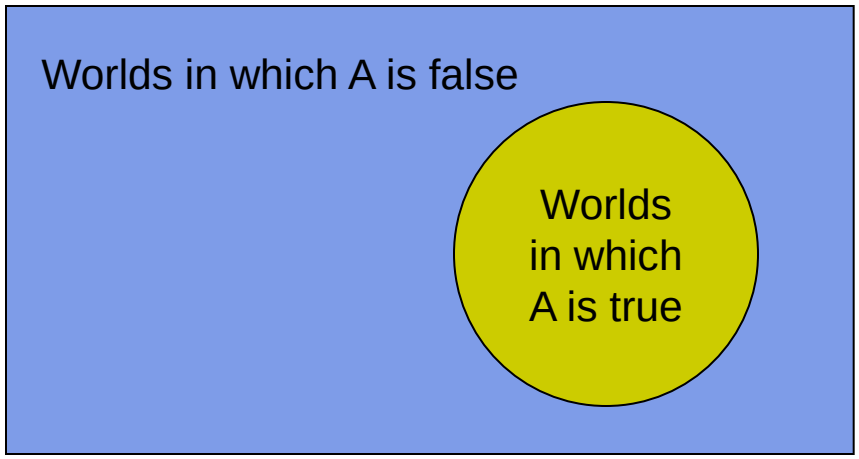
Probability

- A *probability* *P(A)* is a function that maps an event *A* onto the interval *[0, 1]*. *P(A)* is also called the probability measure or probability mass of *A*.

Sample space of all possible worlds.

Call it E

Its area is 1

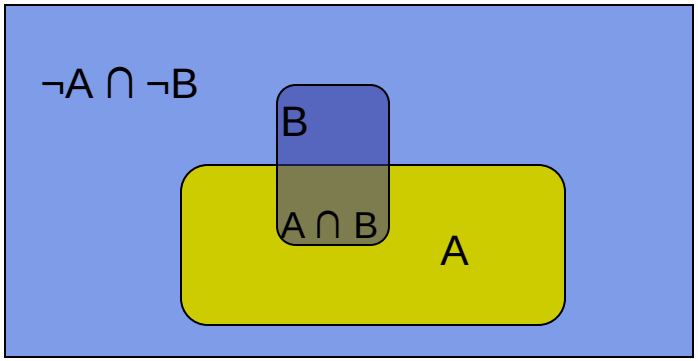


P(A) is the area of the oval

Kolmogorov Axioms

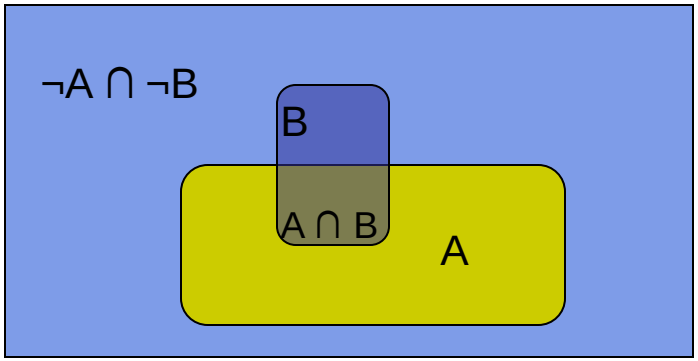
1. All probabilities are non-negative
    1. $0 \leq P(A)$ *for all A*
2. $P(E) = 1$
3. $P(A_1 \cup A_2 \ldots) = P(A_1) + P(A_2) + \ldots.$
    1. If the A_i are pairwise disjoint, $A_i \cap A_j = 0$ *for all I,j*

- All other results about probability derive from these axioms



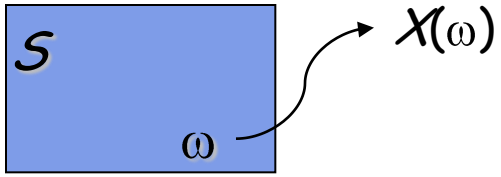
Consequences of Axioms

- $P(\Phi) = 0$.
  - Proof?
- $P(A^C) = 1 - P(A)$
  - Proof?
- $P(A) \leq P(B)$ if A is a subset of B
  - Proof?
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
  - Proof?



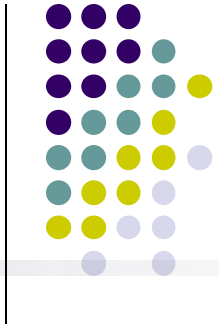
Random Variable

- A *random variable* is a function that associates a unique number with every outcome of an experiment.



- Discrete r.v.:
  - The outcome of a dice-roll: D={1,2,3,4,5,6}
  - Binary event and indicator variable:
    - Seeing a “6" on a toss $\Rightarrow$ X=1, o/w X=0. X_i
    - This describes the true or false outcome a random *event*.
  - Continuous r.v.:
    - The outcome of **observing** the **measured** location of an aircraft X_{obs}

Probability distributions



- For each value that r.v X can take, assign a number in [0,1].
- Like the probability measure defined earlier.
- Suppose X takes values $v_1,\ldots v_n$.
- Then,
  - $P(X= v_1)+\ldots+P(X= v_n)= 1$.
- Intuitively, the probability of X taking value v_i is the frequency of getting outcome represented by v_i

Discrete Distributions

- Bernoulli distribution: Ber(p)

$$P(x) = \begin{cases} 1-p & \text{for } x = 0 \\ p & \text{for } x = 1 \end{cases} \quad \Rightarrow \quad P(x) = p^x (1-p)^{1-x}$$

- Binomial distribution: Bin(n,p)
  - Suppose a coin with head prob. p is tossed n times.
  - What is the probability of getting k heads?
  - How many ways can you get k heads in a sequence of k heads and n-k tails?

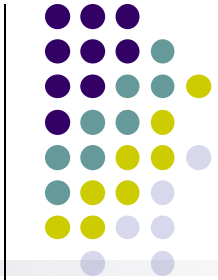
$$\Pr(K = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

More distributions- Multinomial

- Consider a k-sided die.
  - Similar to a coin, but with more possible outcomes.
- A die is tossed n times. What is the probability of getting x_1 ones, x_2 twos…, x_k k's ?
- Let $x=(x_1, x_2, \ldots, x_k)$

$$p(x) = \frac{n!}{x_1! x_2! \cdots x_K!} \theta_1^{x_1} \theta_2^{x_2} \cdots \theta_K^{x_K} = \frac{n!}{x_1! x_2! \cdots x_K!} \theta^x$$

Continuous Prob. Distribution



- A continuous random variable X is defined on a continuous sample space: an interval on the real line, a region in a high dimensional space, etc.
  - X usually corresponds to a real-valued measurements of some property, e.g., length, position, …
  - It is meaningless to talk about the probability of the random variable assuming a particular value --- $P(x) = 0$
  - Instead, we talk about the probability of the random variable assuming a value within a given interval, or half interval, or arbitrary Boolean combination of basic propositions.
    - $P(X \in [x_1, x_2])$
    - $P(X < x) = P(X \in [-\infty, x])$
    - $P(X \in [x_1, x_2] \cup [x_3, x_4])$

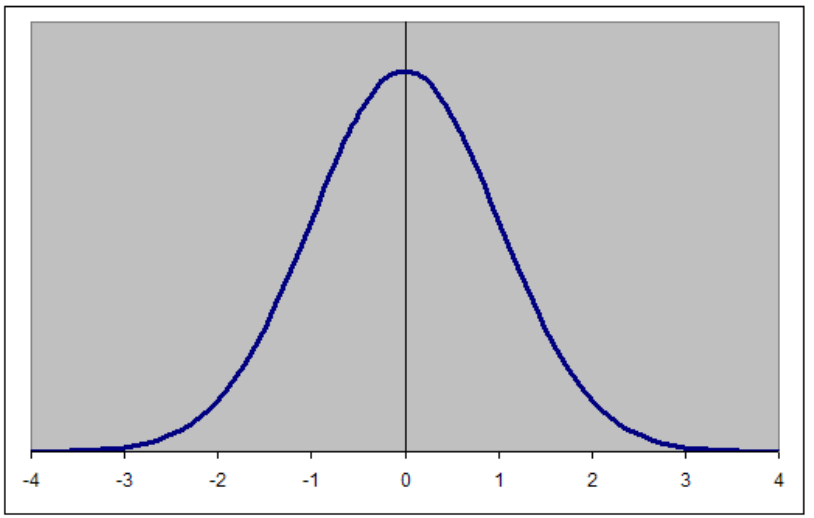
Probability Density

- If the prob. of x falling into $[x, x+dx]$ is given by $p(x)dx$ for dx, then $p(x)$ is called the probability density function over x.
- The probability of the random variable assuming a value within some given interval from x_1 to x_2 is equivalent to the area under the graph of the probability density function between x_1 and x_2.

- Probability mass:

$$P(x \in [x_1, x_2]) = \int_{x_1}^{x_2} p(x)dx,$$

$$\int_{-\infty}^{+\infty} p(x)dx = 1.$$

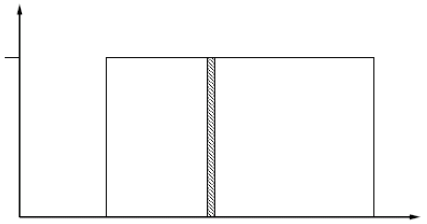


Gaussian Distribution

Continuous Distributions

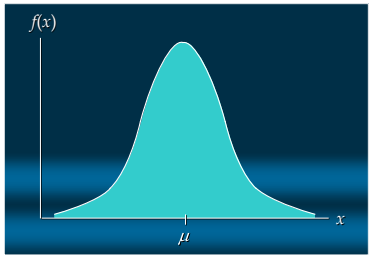
- Uniform Density Function

$$p(x) = 1/(b-a) \quad \text{for } a \le x \le b$$
$$= 0 \quad \text{elsewhere}$$



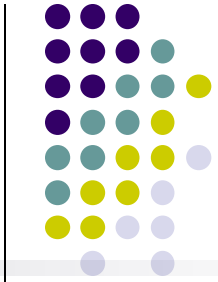
- Normal (Gaussian) Density Function

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$



  - The distribution is symmetric, and is often illustrated as a bell-shaped curve.
  - Two parameters, μ (mean) and σ (standard deviation), determine the location and shape of the distribution.
  - The highest point on the normal curve is at the mean, which is also the median and mode.

Back to RVs - CDF



- Cumulative Distribution Function.
- In a single dice roll, what is the probability of the number rolled being less than 4?
  - P(x<4)=?
  - P(x<4)=P(x=1 OR x=2 OR x=3 OR x=4)
  - But that is the same as P(x=1)+P(x=2)+P(x=3)+ P(x=4).
- A function to represent this quantity is called the Cumulative Distribution Function.
- $F_X(x) = P(X \leq x)$

CDF details

- Definition for a continuous probability function

$$P(x) = P(X < x) = \int_{-\infty}^{x} p(x')dx'$$

- Property of continous CDF:

$$p(x) = \frac{d}{dx} P(x)$$

- Does it have any monotonicity property?

Statistical Characterizations

- **Expectation:** the centre of mass, mean, first moment):

$$E(X) = \begin{cases} \sum_{i \in S} x_i p(x_i) & \text{discrete} \\ \int_{-\infty}^{\infty} x p(x) dx & \text{continuous} \end{cases}$$

  - Sample mean:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$$

- **Variance:** the spread:

$$Var(X) = \begin{cases} \sum_{x \in S} [x_i - E(X)]^2 p(x_i) & \text{discrete} \\ \int_{-\infty}^{\infty} [x - E(X)]^2 p(x) dx & \text{continuous} \end{cases}$$

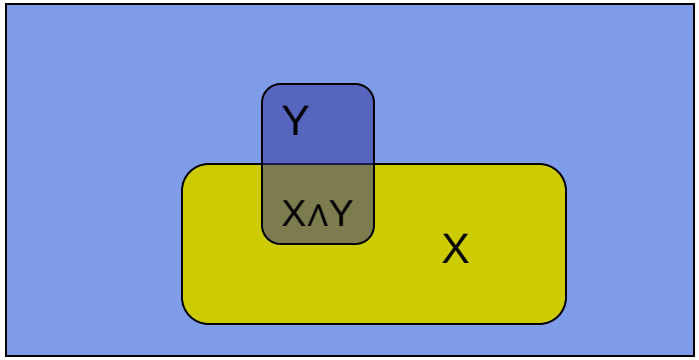
  - Sample variance:

$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu)^2$$

Elementary manipulations of probabilities

- Set probability of multi-valued r.v.
  - $P(\{x=Odd\}) = P(1)+P(3)+P(5) = 1/6+1/6+1/6 = ½$
  - $P(X = x_1 \vee X = x_2, \ldots, \vee X = x_i) = \sum_{j=1}^{i} P(X = x_j)$
- Multi-variant distribution:
  - **Joint probability**: $P(X = true \wedge Y = true)$

    $P(Y \wedge (X = x_1 \vee X = x_2, \ldots, \vee X = x_i)) = \sum_{j=1}^{i} P(Y \wedge X = x_j)$
  - **Marginal Probability:** $P(Y) = \sum_{j \in S} P(Y \wedge X = x_j)$



Joint Probability

- A joint probability distribution for a set of RVs gives the probability of every atomic event (sample point)
  - **P**(*Flu,DrinkBeer*) = a 2 × 2 matrix of values:

	B	¬B
F	0.005	0.02
¬F	0.195	0.78

- **P**(*Flu*) = ?
  - =P(Flu,DrinkBeer)+P(Flu, ¬DrinkBeer) (How?)
  - **= 0.005+0.02**
  - **=0.025**

Conditional Probability

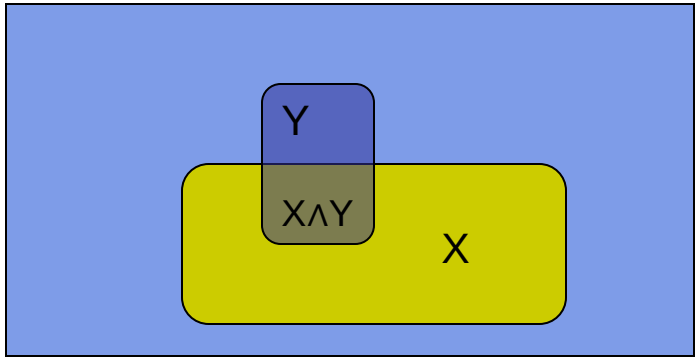
- P(X|Y) = Fraction of worlds in which X is true that also have Y true
  - H = "having a headache"
  - F = "coming down with Flu"
    - P(H)=1/10
    - P(F)=1/40
    - P(H|F)=1/2
  - P(H|F) = fraction of flu-inflicted worlds in which you have a headache

    = P(H∧F)/P(F)
- Definition:

$$P(X \mid Y) = \frac{P(X \wedge Y)}{P(Y)}$$

  - Corollary: The Chain Rule

$$P(X \wedge Y) = P(X \mid Y)P(Y)$$



The Bayes Rule

- $P(Y \mid X)P(X) = P(X \cap Y) = P(X \mid Y).P(Y)$
- Rearrangement gives

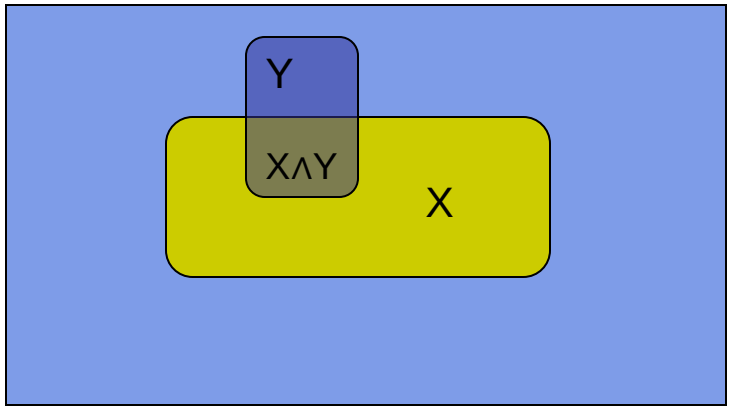
$$P(Y \mid X) = \frac{P(X \mid Y)p(Y)}{P(X)}$$

- This is called Bayes rule

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London,* **53:370-418**

Independence

- Random variables X and Y are said to be independent if:
  - $P(X \cap Y) = P(X)*P(Y)$
- Alternatively, this can be written as
  - $P(X \mid Y) = P(X)$ and
  - $P(Y \mid X) = P(Y)$
- Intuitively, this means that telling you that Y happened, does not make X more or less likely.
- Note: This does not mean X and Y are disjoint!!!

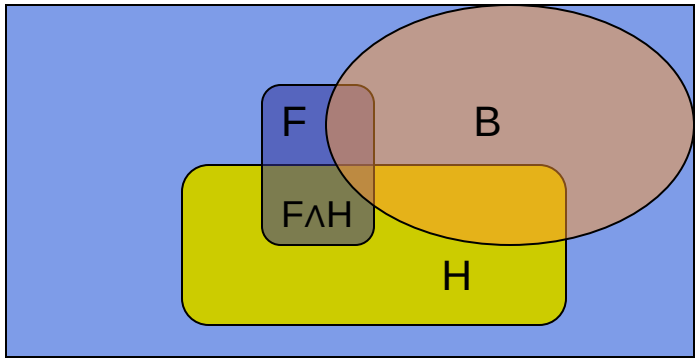


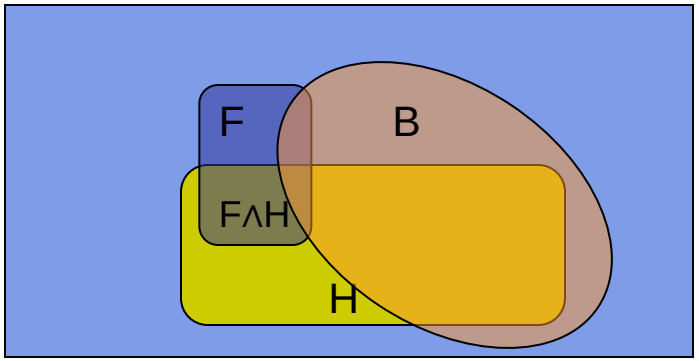
More General Forms of Bayes Rule

- $$P(Y \mid X) = \frac{P(X \mid Y)p(Y)}{P(X \mid Y)p(Y) + P(X \mid \neg Y)p(\neg Y)}$$
- $$P(Y = y_i \mid X) = \frac{P(X \mid Y)p(Y)}{\sum_{i \in S} P(X \mid Y = y_i)p(Y = y_i)}$$
-

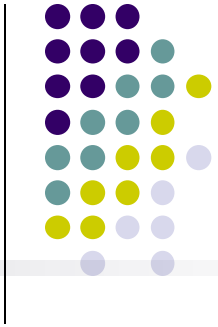
$$P(Y \mid X \wedge Z) = \frac{P(X \mid Y \wedge Z)p(Y \wedge Z)}{P(X \wedge Z)} = \frac{P(X \mid Y \wedge Z)p(Y \wedge Z)}{P(X \mid \neg Y \wedge Z)p(\neg Y \wedge Z) + P(X \mid \neg Y \wedge Z)p(\neg Y \wedge Z)}$$

- P(Flu | Headache ∧ DrankBeer)





Probabilistic Inference



- H = "having a headache"
- F = "coming down with Flu"
  - P(H)=1/10
  - P(F)=1/40
  - P(H|F)=1/2

- One day you wake up with a headache. You come with the following reasoning: "since 50% of flues are associated with headaches, so I must have a 50-50 chance of coming down with flu”

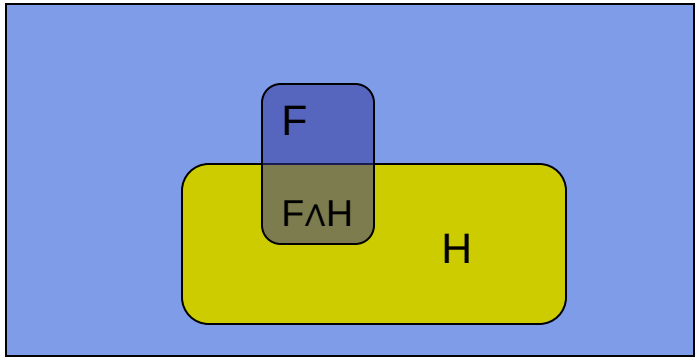
  Is this reasoning correct?

Probabilistic Inference

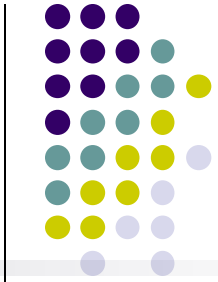
- H = "having a headache"
- F = "coming down with Flu"
  - P(H)=1/10
  - P(F)=1/40
  - P(H|F)=1/2

- The Problem:

P(F|H) = ?



Solution

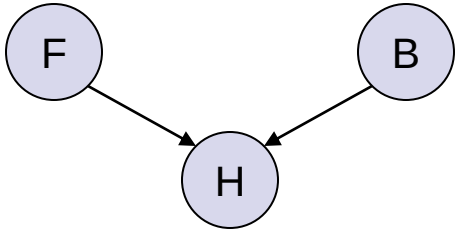


- By Bayes rule
- P(F|H)
  - = P(H|F)P(F) / P(H)
  - =0.5 * 0.025 / 0.1
  - =0.125

- So the probability that you have a flu given that you have a headache is only 0.125 (and not 0.5).
- Also, the probability that you have a flu given that you have a headache is 4 times less than the probability that you have a headache if you are known to have the flu.
  - Why? (Hint: Priors)

Prior Distribution

- Support that our propositions about the possible has a "causal flow"
  - e.g.,

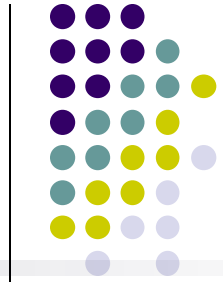


- Prior or unconditional probabilities of propositions

  e.g., P(*Flu =true*) = 0.025 and *P*(*DrinkBeer =true*) = 0.2

  correspond to belief prior to arrival of any (new) evidence

Rules of Independence --- by examples



- P(Virus | DrinkBeer) = **P(Virus)**
  iff Virus is independent of DrinkBeer

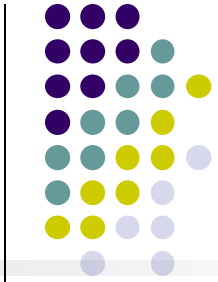
- P(Flu | Virus,DrinkBeer) = **P(Flu|Virus)**
  iff Flu is independent of DrinkBeer, given Virus

- P(Headache | Flu,Virus.DrinkBeer) = **P(Headache|Flu,DrinkBeer)**
  iff Headache is independent of Virus, given Flu and DrinkBeer

Posterior conditional probability

- Conditional or posterior probabilities
  - e.g., $P(Flu|Headache) = 0.178$ incorporates effect of information about the Headache into your probability distribution.
- Representation of conditional distributions:
  - **P**(*Flu*|*Headache*) = 2-element vector of 2-element vectors
- If we know more, e.g., DrinkBeer is also given, then we have
  - **P**(*Flu*|*Headache,DrinkBeer*) = 0.070 **This effect: explaining away!**
  - **P**(*Flu*|*Headache,Flu*) = 1
  - Note how the validity of a certain belief increases or decreases after more evidence arrives, but is not always useful
- New evidence may be irrelevant, allowing simplification, e.g.,
  - **P**(*Flu*|*Headache,SteelersWin*) = **P**(*Flu*|*Headache*)
  - This kind of inference, sanctioned by domain knowledge, is crucial

Conditional independence



- Write out full joint distribution using chain rule:

  P(Headache,Flu,Virus,DrinkBeer)

  = P(Headache | Flu,Virus,DrinkBeer) P(Flu,Virus,DrinkBeer)

  = P(Headache | Flu,Virus,DrinkBeer) P(Flu | Virus,DrinkBeer) P(Virus | DrinkBeer) P(DrinkBeer)

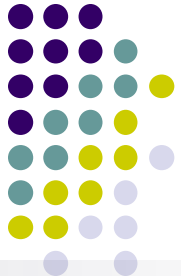
Assume independence and conditional independence in slide 29

= **P(Headache|Flu,DrinkBeer) P(Flu|Virus) P(Virus) P(DrinkBeer)**

I.e., ? independent parameters

- In most cases, the use of conditional independence reduces the size of the representation of the joint distribution from **exponential** in *n* to **linear** in *n*.
- Conditional independence is our most basic and robust form of knowledge about uncertain environments.

Marginal and Conditional Independence



- Recall that for events E (i.e. $X=x$) and H (say, $Y=y$), the conditional probability of E given H, written as $P(E|H)$, is

$$P(E \text{ and } H)/P(H)$$

(= the probability of both E and H are true, given H is true)

- E and H are (marginally) independent if

$$P(E) = P(E|H)$$

(i.e., prob. E is true doesn't depend on whether H is true); or equivalently

$$P(E \text{ and } H)=P(E)P(H).$$

- E and F are *conditionally* independent given H if

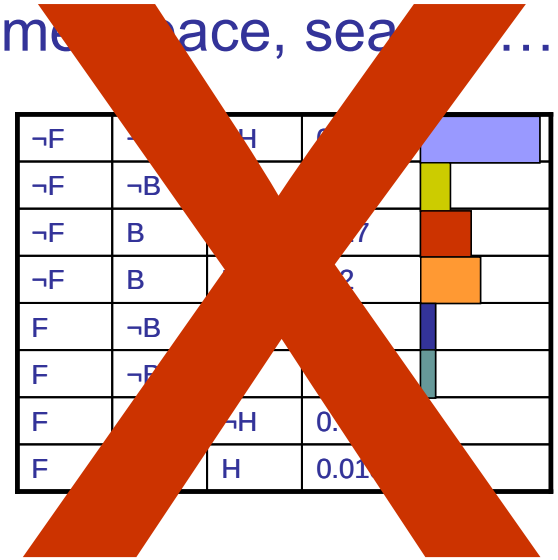
$$P(E|H,F) = P(E|H)$$

or equivalently

$$P(E,F|H) = P(E|H)P(F|H)$$

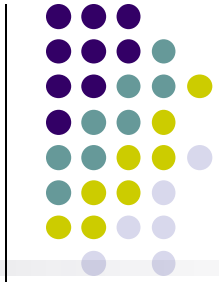
Why knowledge of Independence is useful

- Lower complexity (time, space, search, …)



- Motivates efficient inference for all kinds of queries
- Structured knowledge about the domain
  - easy to learning (both from expert and from data)
  - easy to grow

Information theory



- How do you quantify information?
- Developed by Shannon to find limits on compression, reliable communication and other operations on data.
- Fundamental concept: Entropy

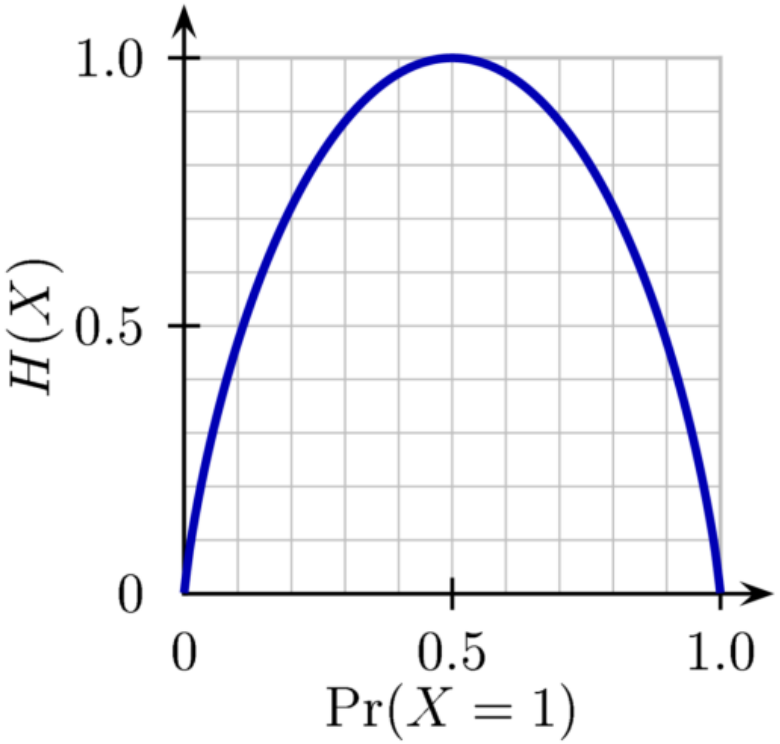
- Entropy – How many bits are needed to convey a message on average?
  - A measure of unpredictability of the message.
  - Messages can be considered to be values of a random variable.

Entropy

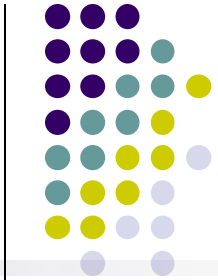
- Usually denoted by H(X) for r.v X

- $H(X) = E(-\log_2(X))$
- $H(X) = -\sum_x p(X=x) \log_2(p(X=x))$

- Entropy for a coin with head probability p
  - -p*log(p)-(1-p)*log(1-p)
  - Is a function of “p”
  - When is it maximum?

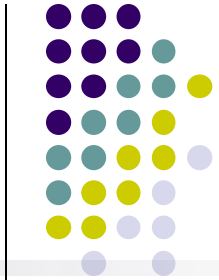


Conditional entropy



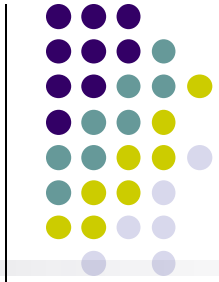
- $H(X \mid Y = y) = -\sum_x p(X=x|y=y) \log_2(p(X=x|y=y))$
  - Y takes a specific value
- Conditional entropy is defined as H(X|Y)
- $H(X|Y) = \sum_y p(Y=y)H(X|y=y)$
- What if $X \perp Y$?
  - Consider what happens to $p(X=x \mid Y=y)$

Mutual information



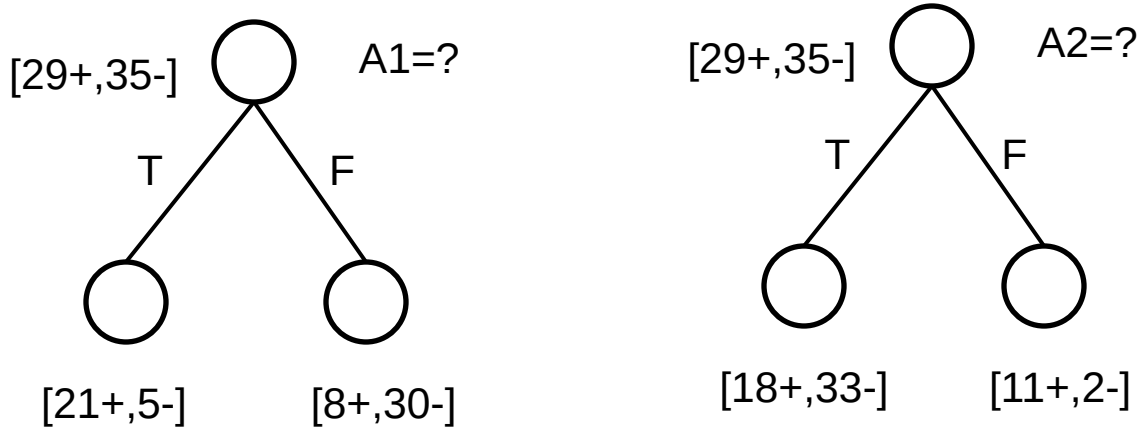
- How much information about X is gained by knowing about Y?
- Alternatively, how much is the unpredictability in X reduced?
  - What is the change in entropy?
- MI(X,Y)=H(X)-H(X|Y) = H(Y)-H(Y|X)
  - Also represented as I(X,Y)
  - =H(X)+H(Y)-H(X,Y)
- I(X,Y) ≥ 0
  - Proof uses Jensen's inequality

Decision trees and entropy



- How to choose what node to split on in a decision tree?
- Class example
  - Decide based on information gain
- Gain(S,A) = Mutual information between attribute A and label Y over the sample set S (Remember S = (X,Y) pairs)
  - How much can you reduce entropy of the label distribution by splitting over attribute A?

Example from class



- Procedure: Find information gain for both split choices
- Choose the one which has larger information gain, i.e, most reduction in entropy
- Gain = Entropy(root) – Weighted Mean(Entropy of children nodes)
  - Weight = Number of points in the node

Example (contd)

- For split by A1

  At root, 29 + and 35 – means

  P(+) = 29/(29+35)=0.45 ; P(-)=35/(29+35) = 0.55

  H(root) = -(0.45*$\log_2$(0.45)+0.55*$\log_2$ (0.55))

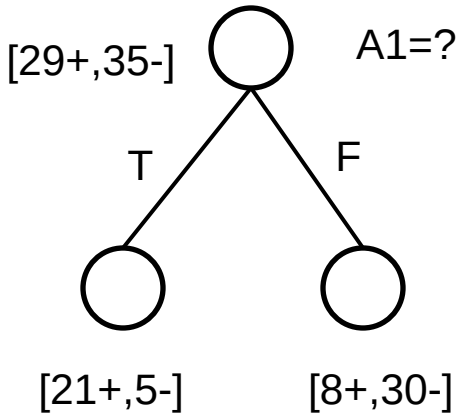
  =0.99

  H(Child 1)= H(21+,5-) = 0.71 (26 points)

  H(Child 2)= H(8+,30-) = 0.74 (38 points)

  Gain(A1)=0.99-[(26/64)*0.71 + (38/64)*0.74)]

  =0.26



Example (contd)

- For split by A2

  At root, 29 + and 35 – means

  P(+) = 29/(29+35) ; P(-)=35/(29+35)

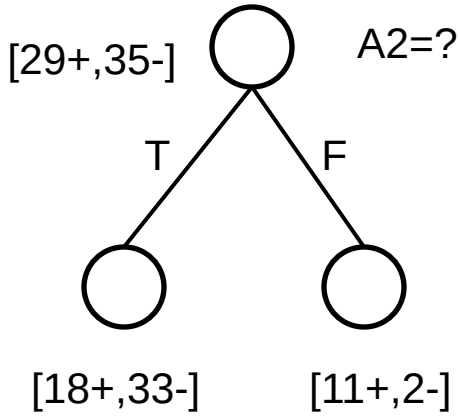
  H(root) = 0.99

  H(Child 1)= H(18+,33-)=0.94 (51 points)

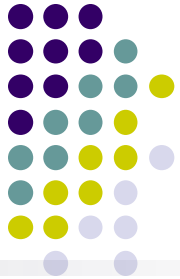
  H(Child 2)= H(11+,2-)=0.62 (13 points)

  Gain(A2)=0.99-[(51/64)*0.94 + (13/64)*0.62)]

  =0.12

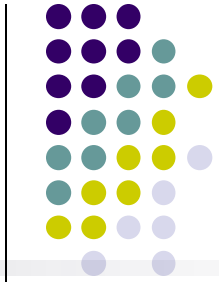


Example



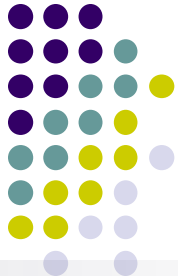
- Gain(A1) = 0.26
- Gain(A2) = 0.12
- Choose to split by attribute A1 since it causes larger reduction in entropy.
- This choice of split is greedy
  - There is no looking ahead, just choose what looks best at this point.
- Easy to implement
  - Hard to guarantee a good result.
- Used in ID3, C4.5 (commonly available decision tree implementations)

Overfitting and decision trees



- We build models to make predictions about labels of unseen data (called test data – attributes X known, label Y unknown) after learning from data with known labels (called training data – both, attributes X and label Y known).
- The objective therefore is “to generalize”.
- The way to do this usually is to reduce your prediction error on training data
  - For example, you construct a Decision Tree using your training data that gives low error if you provide it with examples from your training data and check the predicted label against the known label.
- Sometimes, you can learn your training data too well and lose your ability to generalize.

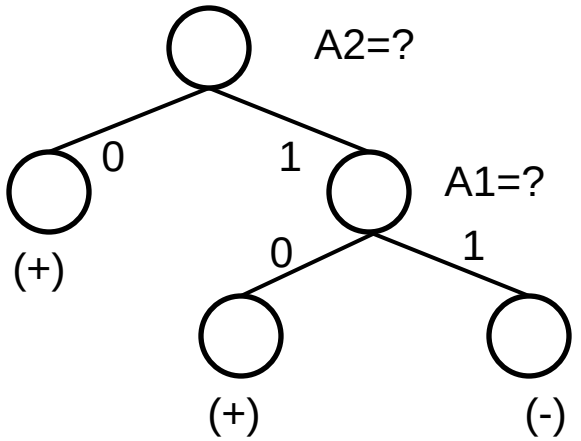
Decision tree overfitting



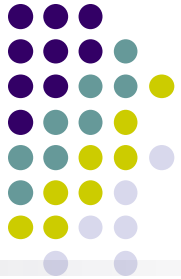
- Suppose you are given a data set with N training examples
  - No two are inconsistent, i.e, there are no examples (X1,Y1) and (X2,Y2) such that X1=X2 but Y1 ≠ Y2
- Can you construct a decision tree that has zero error on training examples?
  - Hint: Consider the decision tree with N leaf nodes.
- Now, what if your test set contains an example not consistent with your training set?
  - Note: This could be caused by your training set being finite.

Example

- X has two attributes binary A1 and A2, each combination is equally likely
  - If X=(0,0) , P(+)=1
  - If X=(0,1) , P(+)=1
  - If X=(1,0) , P(+)=1
  - If X=(1,1) , P(+)=0.9
- Suppose your training set was 1 of (0,0,+), 1 of (0,1,+), 1 of (1,0,+), 1 of (1,1,-)
- Your decision tree would be
- Test data = 100 (1,1) examples (say)
- Error?

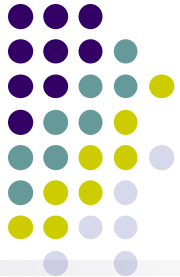


Overfitting



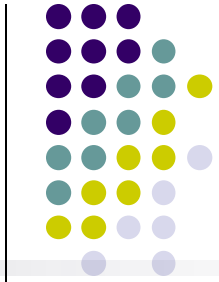
- Often occurs because we get finite training data
  - Sampling from the population is not perfect
- Different algorithms show different degrees of overfitting
- Can be avoided by modifying algorithm suitably.
- In decision trees, a common mechanism is pruning.

Pruning



- Removing nodes from the decision tree.
- A small tree will not give enough prediction accuracy.
- A large tree may cause overfitting.
- When removing nodes
  - How to decide when to stop pruning?
- Many metrics exist
- Two common ways
  - Reduced-error pruning
  - Rule post-pruning

Reduced-error pruning

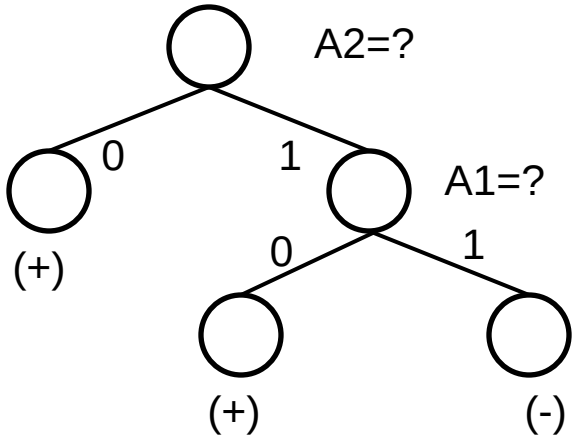


- Keep aside some training data (called a validation set).
  - Not used for training
- Check error of tree on validation data
  - Replace a node by the most popular label at the node
  - Re-check error
  - If node error is reduced, confirm the replacement.
- Practically, look at multiple nodes and prune the one that gives maximum error reduction.

- Simple but not necessarily optimal

Rule post pruning

- Each path to a leaf node in a decision tree is a rule
- Example
  - (A2=1 ^ A1 =1) => (-)
- Prune this rule to
  - A2=1 => (-)
- Check the validation error
  - On original rule
  - On pruned rule



- If error reduced after pruning, modify the decision tree appropriately.